基于全域微观模型的研究前沿主题探测和特征分析*

■ 崔宇红[1] 王飒[1] 高晓巍[2] 杨卉[3] 曹学伟[2]

[1] 北京理工大学图书馆 北京 100081 [2] 中国科协创新战略研究院 北京 100012

[3] 励德爱思唯尔信息技术(北京)有限公司上海分公司 上海 200040

摘要:[**目的/意义**]研究前沿的准确判断是国家宏观层面的战略需求,文献计量学作为一种定量研究方法广泛应用于科学主题探测和研究前沿识别中。[**方法/过程**]梳理研究前沿主题探测的发展历程和方法模型,引入全域微观模型的概念,详细介绍 SciVal 模块采用的主题创建方法,包括直接引用文献聚类、关键词主题命名和研究前沿遴选的主题显著性算法,并对 SciVal 创建的 9.6 万个主题和遴选出的前 1% 的研究前沿主题的特征进行实证分析。[**结果/结论**]全域微观模型可以同时一次识别整个科学领域的所有主题,但不同学科在研究前沿上表现存在差异,不能把主题显著性简单等同为重要性;主题论文数量与主题排名之间存在中度相关性;自动抽取的关键词术语从学科领域层和独特性上命名和描述主题;石墨烯相关前沿主题的演变趋势分析可以用于发现关键节点和新兴主题。

关键词: 主题探测 研究前沿 全域微观模型 SciVal 主题显著性

分类号: G301

DOI: 10.13266/j.issn.0252-3116.2018.15.009

一个学科领域的研究前沿是最能代表该学科的发展趋势、制约该学科当前发展的重大关键性问题。从宏观层面的战略需求上看,研究前沿的准确判断会影响一个国家科学、技术和创新发展的政策导向。日本、欧盟、美国和加拿大等为了成为全球科学技术的领导者并保持其科技强国的地位,从 2006 年起就开始将研究前沿作为首要研究课题,成立面向创新前沿的研究机构和专项基金,支持前沿性研究[1]。为了加快实施国家的创新驱动发展战略,我国国务院 2016 年 5 月印发的《国家创新驱动发展战略纲要》中,进一步指出要"加强面向国家战略需求的基础前沿和高技术研究"和"面向科学前沿加强原始创新"等任务,从国家战略高度指明了科学前沿研究的重要性与紧迫性。文献计量学作为一种定量研究方法广泛用于科学主题探测和研究前沿识别中。主题探测和前沿识别面临诸多问题,例如,科学中有多少个主题?所有主题应该在整个科学领域被一次识别,还是可以按照特定需要只在比较小的领域上被识别?哪种方法提供最精准的主题描述?研究前沿具有哪些特征?如何有效解读和展现研究前沿?近年来,基于整个科学领域所有文献直接引用关系的全域微观(Global-micro)模型被引入并应用于主题创建和前沿识别研究中,2017 年 10 月,SciVal 采纳此模型推出了主题显著性模块,为上述问题提出了一种新的解决方案。

1 研究前沿主题探测综述

科学主题探测和研究前沿的识别始终吸引着科学家的兴趣。早在 1955 年,加菲尔德就在广为人知的《科学引文索引》中指出,科学文献的引用链接分析可以跟踪新兴思想和发现科学的新兴领域[2]。1965 年,普赖斯利用大量的引文数据定义了他所描述的"科学研究前沿",即某些卓越科学家在最前沿领域进行的领先研究,并从出版物的密度以及不同时期的活跃度对研究前沿进行了测度[3]。1970 年,社会学家库恩明确提到可以用加菲尔德的引文数据来识别研究社区和描绘科学革命的范式[4]。

* 本文系中国科协创新战略研究院研究课题"基于大数据和科学计量方法的科技前沿探测研究"成果之一。

作者简介:崔宇红(ORCID:0000-0002-5215-7726),研究馆员,博士,E-mail:cuiyh@bit.edu.cn;王飒(ORCID:0000-0003-4665-5419),副研究馆员,博士;高晓巍(ORCID:0000-0001-6811-8759),博士;杨卉(ORCID:0000-0002-5734-415X),咨询主管;曹学伟(ORCID:0000-0001-7387-214X),博士后。

收稿日期:2017-12-08 修回日期:2018-04-07 本文起止页码:75-82 本文责任编辑:王传清

在前人的理论基础上,不同时期的文献计量学家采用直接引用(Direct Citation,DC)、共被引(Co-Citation,CC)和文献耦合(Bibliographic Coupling,BC)3种文献引用关系(见表1),基于不同的数据源、文献规模和聚类算法,开展了基于引文模型方法的探索和实证研究,从一个侧面反映了科学图谱理论、数据可视化技术和计算机信息处理能力的演变过程。作为ISI的创始人,尽管加菲尔德早在1964年就提出直接引用分析可以用于构建发现科学突破的历史图谱[5],但由于直接引用会产生大量的计算需求,受早期计算处理能力的限制并没有得到广泛使用。1965年,凯斯勒分析了《物理学评论》上334篇论文的文献耦合关系,这无疑是当时最大规模的文献聚类研究[6]。

表1 标志性研究成果及其采用的方法模型

作者/产品	发表时间(年)	数据来源	文献数量(条)	引用关系	聚类算法
Kessler	1965	Physical Review	312	BC	Single-link
Small & Griffith	1974	ISI	1 832	CC	Single-link
Small	1999	ISI	164 612	CC	Single-link
ESI Research Fronts	2001	ISI	*	CC	Single-link
Klavans & Boyack	2006	ISI	731 289	CC/BC	VxOrd
Boyack	2009	ISI	997 775	BC	VxOrd
Klavans & Boyack	2010	Scopus	2 080 000	CC	DrL/OpenOrd
Waltman & van Eck	2012	ISI	10 200 000	DC	Smart Local Moving
Boyack & Klavans	2014	Scopus	20 431 588	DC	Smart Local Moving
SciVal ToP in Science	2017	Scopus	~70 000 000	DC	VOS

第一个广为使用的研究前沿探测模型是1974年由ISI首席科学家斯莫和格里菲斯提出的,他们基于ISI的1 832篇高被引论文的共被引分析和Single-link聚类算法,展现了科学的完整结构图谱[7],应用于科学新兴领域的跟踪和预测研究中[8],并一直持续使用至今。从1974年到2010年,从斯莫到克拉万斯和博雅克,尽管文献规模从千篇增长到百万级[9],聚类算法从单点链接到VxOrd和DrL/OpenOrd[10],但是共被引分析和ISI数据库的组合几乎没有变化,绝大多数的研究者通过限定学科、期刊或者术语在局域数据集合上创建文献聚类,实现特定研究的前沿主题探测和识别。2001年,ESI采用共被引分析对Web of Science的高被引论文聚类,一次性识别并动态生成22个学科领域的近万个研究前沿,国内学者基于ESI前沿主题的数据进一步开展了纳米前沿领域图谱[11]、生物科学前沿演进时序[12]、量子失协领域关键研究路径[13]等实证分析。

近年来研究前沿探测的挑战是如何精确地构建整个科学领域上更加精细的主题识别模型框架。继共被引分析被广泛应用后,克拉万斯和博雅克在2009年和2010年分别在ISI[14]和Scopus[15]的百万级文献规模上,尝试采用文献耦合和共被引分析与不同的聚类算法的组配,2011年又提出全域微观模型的概念并不断完善[16]。2012年,荷兰莱顿大学的沃特曼和凡埃克提出了第一个基于直接引用模型构建整个科学领域的论文级分类体系的新方法[17],他们证明基于直接引用和Smart Local Moving聚类算法可以将千万级的ISI论文精确划分为不同的主题,同时这种方法简单透明,对计算设备的性能要求不高。2014年,博雅克和克拉万斯采用沃特曼等的方法处理了超过2 000万的Scopus文献数据[18],之后两位研究者比较和评估基于3种数据引用类型在构建研究主题科学图谱的效果,发现直接引用要比文献耦合或共被引分析能够更精确地绘制微观研究问题层级的知识分类体系,更好地发现新兴交叉学科,理解整个科学领域的发展趋势和演化动力[19]。

研究前沿探测包括主题创建和前沿遴选两个阶段。2017年10月,爱斯维尔SciVal在主题创建过程中采纳全域微观模型,对Scopus中从1996年到2016年以来所有科学领域的7 000万论文和参考文献进行聚类,识别形成近9.6万个研究主题。研究表明,研究前沿最普遍的特征是高关注度和新颖性[20],如ESI研究前沿用高被引作为高关注度的计算依据,而《2017研究前沿》报告则按照核心论文出版年排序,找出"最年轻"的研究前沿用于深入的解读分析[21]。与之不同的是,SciVal基于近2年论文的引用、浏览和期刊质量指标,综合计算每个主题的显著性百分位数,可以看到,显著性百分位数具有高关注度和新颖性两个特征,因此,利用SciVal的主题显著性数据遴选出位于所有学科领域上最大、最热的研究前沿并验证其效果是本文

研究的主要目的。

本研究第二部分介绍全域微观模型概念，以及SciVal应用此模型中的主题创建、关键词主题命名和研究前沿遴选的方法。第三部分以SciVal平台的近9.6万个主题及前1%研究前沿为对象，分析学科分布特征，验证主题论文数量与主题排名之间的关系，如何用关键词描述主题，并以石墨烯研究为例，展示主题演变趋势。最后讨论全域微观模型的优势和存在的问题，指出下一步研究内容。

2 模型和方法

下面介绍主题创建中采用的全域微观概念模型、直接引用主题聚类模型和关键词主题命名方法，以及从创建的所有主题中遴选研究前沿的主题显著性计算方法。

2.1 全域微观概念模型

全域是指从整个数据库所有文献数据而不是子集来构建数据集合。相比之下，局部模型是在一个文献子集上完成的，由于研究者无法获得数据库中所有数据的访问权限，只能下载部分数据到本地，例如指定物理学学科分类下的所有文献，选择信息科学的若干种期刊，或者检索“石墨烯”术语的高被引论文等。研究表明，全域模型比局部模型在精确性和召回率上要高，更适合于描述和发现那些不能提前预知的突发主题[22]。相关的科学和技术领域可能包括发现的途径，如物理学的新发现可能来自于化学，或者来自计算机科学，或者来自于仪器技术的发展。因此，有必要产生一个涵盖尽可能多的科学和技术文献的模型。

在层次化的科学分类体系中，以前的研究主要在领域、学科或专业方向层次上聚合文献。领域位于树状分类的最顶端，其数量在几个到几十个之间，如ESI的22个学科领域或者Scopus的27个学科领域，一个领域每年约有几十万条文献。学科位于领域的下一级，每年的文献数量从几百到几千，一般是基于期刊聚类，学科经常与Web of Science或者Scopus的学科目录等同。专业方向级的分析经常基于来自术语检索的结果产生的文献样本。

主题由具有同样研究基础的一组文章，在研究问题级或者微观层上聚合文献，处于科学分类体系的底端。研究问题是研究者实际从事的细节问题，例如，虽然“染料敏化太阳能电池”被视为一个专业方向，但“染料敏化太阳能电池的反电极材料研究”则是一个研究问题。主题规模可大可小，每年的文献数量从几篇到几百篇不等。

简单地说，全域微观模型是将所有学科领域上的文献聚合到研究问题级的一种精细的科学分类体系。

2.2 主题聚类模型

主题创建包括用直接引用形成文献簇和将文献簇聚类成不同主题两个步骤。直接引用文献簇创建在概念和实践上都相对简单，图1是一个简化示意图。当处理大规模文献集合时，为减少计算资源，应尽可能减少有关联的文献对的数量，因此采用不考虑引用方向的直接引用，即只要文献 引用了j或者j引用了i，那么$C_{ij}=1$，否则$C_{ij}=0$。

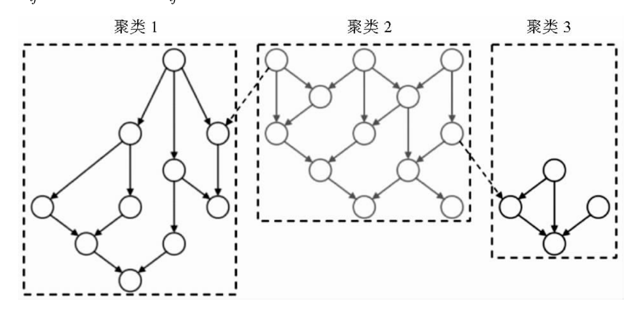


图1 直接引用聚类模型示意[23]

通过直接引用链接创建文献簇后，对直接引用文献簇采用沃特曼和凡埃克开发的VOS方法进行聚类。VOS算法使用了模块性聚类变量最大限度地提高簇到簇之间的相似性的准确度，即用簇的相似性作为权重，越相似的簇权重越高，越低相似的簇权重越低[24]。为计算论文的关联度，每个链接要用施引文献的参考文献数进行标准化，并计算所有矩阵的K50(修正余弦)值。同时为了减少计算规模，将每篇文章的链接上限设置为15(最高K50值)，然后将论文集和过滤后的链接输入到VOS聚类编码中。VOS算法的实现可以免费获得。

SciVal依据直接引用关系创建的研究主题一旦产生后就永久存在，每年会产生少量新的主题，后来的文献根据引用关系增加到主题中，最新文献越多表明主题越新，旧主题不会消失，但可能处于休眠状态。

2.3 主题命名方法

通过上述步骤创建的主题一般由几十到几千篇文献组成，需要用计算机自动抽取关键词或短语来命名主题。主题命名方法综合使用了爱思唯尔的指纹技术(Elsevier Fingerprint Technology, EFT)和特殊短语，通过3步过程来创建主题名称：①应用自然语言处理技术挖掘主题中论文的标题和摘要信息。②用一组词语与所有主要学科的叙词表进行匹配得到概念术语。爱

思维尔集成了若干个通用和专业领域主题词表，如医学主题词表 MeSH、天文学主题词表（Unified Astronomy Thesaurus，UAT）等。③每一个文档基于逆文档频率选择独特的关键词，减少在文档集的高频词的权重，并增加很少出现单词的重要性。按照与最高词频术语的出现频次的比值，每个关键词被给出 0 和 1 之间的相关性值，相关性值为 1 的表示最频繁出现的关键词。

在实际使用中，系统会自动给出 3 个术语来命名每个主题。前两个使用 EFT 生成，一般选择高频词，提供对主题在研究领域或者专业方向高层次上的描述。第三个选择关于此主题的特殊短语，是对主题在研究问题层次上作更具体的描述。例如，一个被命名为“Graphene; Energy storage; Graphene fibers”的主题，研究方向涉及“石墨烯和能源存储”，具体研究内容是“石墨烯纤维”，因此这个研究主题可以描述为“能源存储中的石墨烯纤维”。

2.4　主题显著性计算

主题显著性是一个测度主题的可见度和发展势头的指标[25]，它综合考虑了最近引用数量、最近浏览数量和期刊 Citescore3 个参数，对每个主题 j 在第 n 年的显著性 P_j，计算公式如下，

$$P_j = 0.495(C_j - mean(C_j))/stdev(C_j) + 0.391(V_j - mean(V_j))/stdev(V_j) + 0.1149(CS_j - mean(CS_j))/stdev(CS_j) \quad \text{公式(1)}$$

这里，C_j 是主题 j 中在第 n 年和 $n-1$ 年发表论文的引用量，V_j 是主题 j 中在第 n 年和 $n-1$ 年发表论文的 Scopus 浏览量，CS_j 是主题 j 中在第 n 年发表论文的平均 CiteScore，其中原始数据经过了对数转换，即公式 2：

$$C_j = \ln(C_j + 1), V_j = \ln(V_j + 1), CS_j = \ln(CS_j + 1) \quad \text{公式(2)}$$

显著性计算是用标准化分数消除 3 个指标之间的量纲差异，再对每个主题近两年论文的引用数量、浏览数量、期刊评价指数与平均值的离散程度加权求和。因此，显著性数值越高，表示越来越多的研究者正在关注这个主题，也说明这个主题的增长势头越猛。实际使用中，SciVal 根据主题的显著性数值排序，计算每个主题的百分位数指标。

3　结果分析

3.1　研究前沿的遴选

爱思维尔在 2017 年 10 月推出的新版 SciVal 中，用主题显著性代替了原来的竞争力分析，基于 Scopus 数据库约 7000 万条文献和 10 亿个引用链接采用前述的模型方法进行主题聚类，共得到整个科学领域的近 9.6 万个主题，并给出每个主题的显著性百分位数，表 2 列出显著性百分位数排在前 10 位的研究前沿主题及指标。

表 2　排名前 10 位的显著性主题及指标

主题排名	主题编号	关键词命名	论文数量（篇）	引用数量（次）	浏览数量（次）	Citescore	显著性百分位数
1	T20	Perovskite; Solar cells; methylammonium lead	3 872	33 690	84 002	7.35	100.000
2	T63	Molybdenum compounds; Monolayers; dichalcogenides TMDs	3 808	12 739	30 524	5.96	99.999
3	T456	Genome; RNA, Guide; effector nucleases	2 904	13 321	22 516	5.28	99.998
4	T6	Electrolytic capacitors; Capacitance; asymmetric supercapacitors	4 065	10 557	33 524	4.59	99.997
5	T0	Solar cells; Heterojunctions; organic photovoltaics	4 564	10 837	22 836	6.02	99.996
6	T2050	Sulfur; Electric batteries; lithium polysulfides	1 862	6 699	31 445	6.76	99.995
7	T1727	Electric batteries; Lithium compounds; batteries SIBs	1 902	7 479	26 383	6.85	99.994
8	T3007	Viruses; Infection; ZIKV infections	1 564	11 372	23 321	3.76	99.993
9	T350	Electrolytic reduction; Electrocatalysts; non-precious metal	2 577	7 908	22 102	6.05	99.992
10	T403	Immunotherapy; Melanoma; immune-related adverse	2 290	18 796	9 953	3.81	99.991

百分位数指标作为一种相对指标，近年来在卓越绩效评价中被广泛应用[26]。实际应用中往往根据需求选择合适的百分位分数作为阈值，例如 ESI 的高被引论文和研究前沿都是基于各学科论文引用数量前 1% 的阈值，教育部学科评估将 ESI 高被引论文扩展到前 3%，《2017 研究前沿》报告则在 ESI 研究前沿的基础上进一步提取前 10% 的最具引文影响力研究前沿[21]。本研究拟从 9.6 万个研究主题中遴选出所有学科领域的最新和最热的研究问题，然后再划分到相应的学科领域进行解读分析，以展示当前科技前沿的最新进展。考虑研究前沿的发布需求和解读工作量，以及应保证尽可能涵盖不同的学科领域，经实验，设定主题显著性百分位数阈值为 99%，即前 1% 的主题为研究前沿，共得到 963 个研究前沿主题，涵盖了除艺术

与人文、兽医、健康学、多学科之外的23个学科领域。

3.2 主题学科分布

按照每个主题中论文所属最多的学科的原则，将每个主题的学科归属映射到Scopus的27个学科类目上，表3统计不同学科所有主题和研究前沿主题的数量，研究前沿主题占所有主题的百分比，以及研究前沿主题的相对强度。相对强度是学科研究前沿数量与最大的学科研究前沿数量之间的比值，例如，医学研究前沿主题数量最多为188个，医学研究前沿主题的相对强度为1，化学排在第二位为171个，则化学研究前沿主题的相对强度为0.91。

表3 按Scopus的27个学科分类的主题分布统计

Scopus 学科领域	所有主题数量(篇)	研究前沿主题数量(个)	研究前沿主题占所有主题百分比(%)	研究前沿主题的相对强度
医学	22 039	188	0.85	1.00
工程	12 259	75	0.61	0.40
社会科学	8 995	4	0.04	0.02
农业与生物科学	6 888	36	0.52	0.19
计算机科学	6 038	25	0.41	0.13
物理学和天文学	5 259	71	1.35	0.38
化学	5 247	171	3.26	0.91
艺术与人文	4 525	0	0.00	0.00
生物化学、遗传学和分子生物学	3 811	67	1.76	0.36
材料科学	3 046	126	4.14	0.67
地球与行星科学	2 726	18	0.66	0.10
数学	2 408	3	0.12	0.02
环境科学	2 136	51	2.39	0.27
药理学，毒理学和制药学	1 773	13	0.73	0.07
商业、管理和会计	1 376	8	0.58	0.04
经济学、计量经济学和金融	1 180	2	0.17	0.01
心理学	1 073	3	0.28	0.02
能源学	1 024	33	3.22	0.18
化学工程	947	29	3.06	0.15
免疫和微生物学	902	24	2.66	0.13
兽医	550	0	0.00	0.00
神经系统科学	493	10	2.03	0.05
护理学	394	4	1.02	0.02
牙科	244	1	0.41	0.01
健康学	206	0	0.00	0.00
决策科学	131	1	0.76	0.01
多学科	99	0	0.00	0.00

不同学科在研究主题和研究前沿上的表现存在较大差异。图2是用学科研究前沿主题的相对强度和占所有主题百分比两个指标构建的学科分析矩阵。27个学科可以分为4个集群：①医学（标识O），无论在所有主题还是研究前沿主题中的数量都最高，研究前沿占比为0.85%，略低于1%的预期值；②化学和材料科学（标识×），研究前沿主题数量仅次于医学，相对强度分别为0.91和0.67，但研究前沿占比显著高于1%的预期值，分别为3.26%和4.14%；③能源学、化学工程、免疫和微生物学、环境科学、神经系统科学、生物化学遗传学和分子生物学、物理学和天文学7个学科（标识+），研究前沿主题数量较少，相对强度低于0.4，但研究前沿占比高于1%的预期值；④包括工程、计算机科学在内的17个学科（标识□），研究前沿主题数量和占比上都比较低，特别是艺术和人文、兽医、健康学、多学科等的研究前沿数量甚至为0。

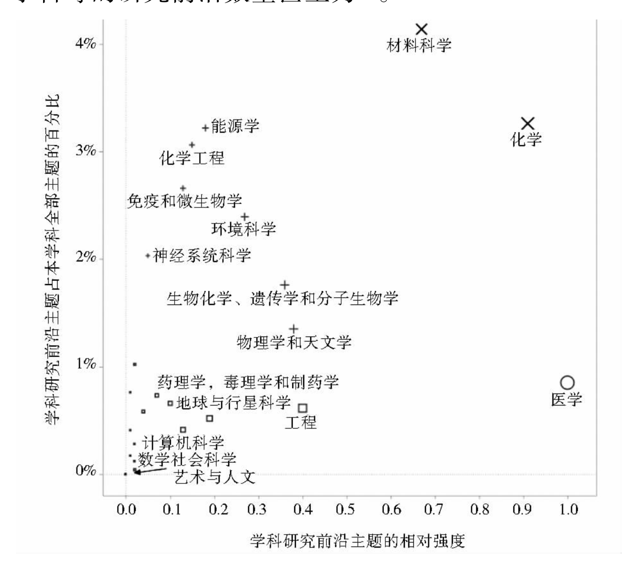


图2 研究前沿主题的学科分析矩阵

3.3 主题规模与排名关系

统计95 769个主题的论文数量（在2012－2016年的5年时间窗内），最多的有4 574篇，最少的仅有1篇，中位数为56篇。相比而言，963个研究前沿主题的论文数量，最少的有122篇，中位数为1 119篇。图3是论文数量与主题排名之间的关系图（坐标轴取对数值），论文数量与主题排名呈中度正相关（$R^2=0.692$，$p<0.0001$），即主题论文数量越多，显著性指标越高，主题越可能排在前列。

对图3中论文数量很少但排名靠前的异常值应特别关注。例如，主题T67927仅有35篇文章，排在第984位（接近前1%的研究前沿），主题关键词为诊断、血液、计算机显微镜。引用主要来自一篇高被引文章，是美国癌症协会发布的2016年癌症统计年报[27]，被引用4 242次（截至2017年10月30日）。类似文献一般都会被高引用，但不能认为是本主题的核心论文。

在另一个例子中，主题 T67378 仅有 122 篇文章，排在第 440 位（位于前 1% 的研究前沿），主题关键词为疾病、卫生服务、糖尿病，有多篇高被引论文，可被认为是本主题的核心论文。这两个例子表明对于导致主题高被引的情况还需进一步的详细考察。

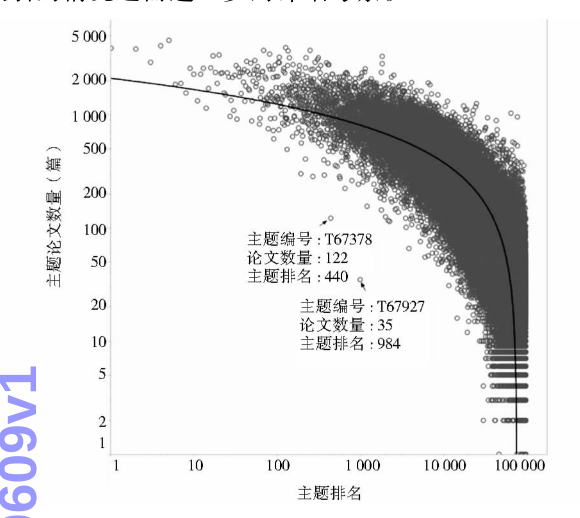


图 3　研究主题的论文数量与排名关系

3.4　关键词主题命名

根据主题命名规则，前两个术语是从题名和摘要抽取出用来描述研究领域或方向层级的关键词，第三个术语则用特殊短语描述主题的专有性。对前 1% 的研究前沿主题的前两个关键词词频统计结果显示，肿瘤、光催化、石墨烯、锂离子、细胞凋亡、碳纳米管、DNA、水合物、蛋白质、太阳能电池、生物燃料、催化剂、能源管理、水凝胶、电催化、宏基因组学、纳米粒子、磷酸盐的出现频次排在前列，均在 10 次以上。这些术语主要涉及医学、化学、材料科学、生命科学、能源学、环境科学、物理、工程等，这也与前述的研究前沿的学科分布态势一致，描述了发展势头迅猛的前沿学科。

为了验证系统自动给出的关键词术语在描述和命名主题中的效果，以“Graphere”为检索词查询全科学领域所有主题的关键词集合，得到 19 个位于前 1% 的与石墨烯相关的研究前沿，分属于化学、材料、物理和工程 4 个学科领域，依据这 3 个关键词初步命名主题。经咨询石墨烯领域专家，认为多数关键词术语从宏观的研究方向和微观内容的独特性上描述主题，可以较好地帮助专业人员快速理解和初步判断一个主题研究的内容是什么。但是也存在关键词不够精确的现象，如主题 T31540，其 3 个关键词为“电解电容器、石墨烯、面积比电容”，专家认为术语“面积比电容”过于狭窄很少使用，这就需要通过进一步解读主题的核心论文，人工给出更精确的描述“石墨烯超级电容器”。

表 4　石墨烯研究前沿主题描述

主题编号	主题排名	主题命名	Scopus 学科目录	论文数量（篇）
T235	48	氧化石墨烯	一般化学	2 850
T6651	50	石墨烯药物释放	一般材料科学	1 326
T1072	71	石墨烯的化学气相沉积法合成	一般材料科学	2 167
T16939	78	能源储存中的石墨烯纤维	一般材料科学	752
T740	79	石墨烯等离子体	原子与分子物理和光学	2 713
T18168	90	石墨烯气凝胶	一般材料科学	723
T6784	95	氧化石墨烯（GO）复合材料的光催化性能	一般材料科学	1 317
T8319	157	石墨烯硅烯纳米带的电子输运性质	凝聚态物理学	1 199
T15753	197	石墨烯液相剥离	一般化学	717
T15956	254	石墨烯锂离子电池	一般材料科学	633
T17039	287	石墨烯吸附的拟二阶模型	一般化学	728
T441	398	锯齿边缘石墨烯的电子输运性质	凝聚态物理学	1 952
T31540	407	石墨烯超级电容器	一般材料科学	374
T17638	482	石墨烯气体传感器	电子电气工程	636
T3084	548	石墨烯热导率的非平衡态分子动力学	凝聚态物理学	1 119
T18995	604	石墨烯薄膜太阳能电池	一般材料科学	662
T2164	759	单层与扭曲双层石墨烯性质	凝聚态物理学	1 277
T6871	815	石墨烯晶体管	电子电气工程	1 134
T17045	894	应变石墨烯	凝聚态物理学	640

3.5　相关主题发展趋势比较：以石墨烯为例

研究前沿是对 1996 – 2016 年间的文献进行聚类形成，并将更新的最近论文依据直接引用关系分配到现有主题中。因此，统计多个相关主题在不同时间的论文分布不仅可以展示一个研究问题的发展变化过程，而且可以发现某个研究领域或研究方向上的关键节点和新兴趋势。因此，统计多个相关主题的论文时间分布，不仅可以展示一个研究问题的发展变化过程，

而且可以发现某个研究领域或研究方向上的关键节点和新兴趋势。

石墨烯排名最前的 6 个前沿研究在 1996 - 2016 年期间论文发表数量的变化趋势见图 5。2004 年英国曼彻斯特大学的研究者通过简单方法剥离出单层石墨烯，导致从 2006 年开始石墨烯研究论文数量明显增加，特别是 2010 年诺贝尔物理学奖的获得极大加速了石墨烯的研究，成为当前最热门的研究领域之一。相比之下，氧化石墨烯和石墨烯等离子体两个研究前沿主题的论文数量近两年保持在 600 篇左右，但石墨烯等离子体的论文数量增长势头更猛，2013 年超过了石墨烯合成，2015 年又超过了氧化石墨烯，成为石墨烯领域当前最受关注的研究主题。石墨烯药物释放类论文近两年保持了较高的数量增长，2016 年论文达到约 400 篇。相比之下，能源储存中的石墨烯纤维和石墨烯气凝胶两个主题是材料科学与能源环境领域的跨学科应用，因此尽管论文数量不高，但依然获得了很高的显著性排名。

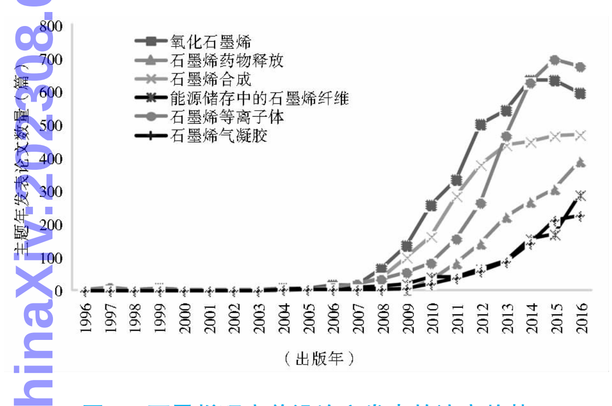


图 5　石墨烯研究前沿论文发表的演变趋势（1996 - 2016 年）

4　思考和展望

本研究从理论和实证角度展示了全域微观模型在创建、识别、遴选和描述整个科学领域的研究前沿上的方法和过程。很多主题本身具有跨学科特征，例如，将石墨烯相关的研究主题中的论文映射到 Scopus 的学科分类体系上，会发现同时覆盖材料科学、化学、物理、工程、化学工程、能源、生物化学、环境科学等多个学科，这也反映当前石墨烯研究的现状。随着跨学科研究越来越成为常态，人们倾向于认为新兴前沿问题或者重大科学突破往往会产生在学科的交叉和边缘地带，SciVal 推出的显著性主题由于事先不限定检索，容易识别跨学科主题，这为研究人员和决策管理者探测新兴研究前沿、制定优先发展方向和分配基金项目等方面提供了一种有效的工具。

需要注意的是，由于不同学科本身存在差异性，例如，计算机科学领域主题显著性会明显低于材料和医学领域，而社会科学和人文艺术等软科学领域与硬科学领域的显著性更不能直接比较。因此，主题显著性不能简单等同于重要性、创新性、新颖性或者热点，一个在全域中显著性较低的主题可能对本领域仍是很重要的，实际应用中要根据不同的识别目的 - - 破性研究还是有技术应用的潜力研究，是新兴前沿还是公众关注的热点，引入更多的数据源并设计相应的遴选指标和方法。

主题显著性排名与主题论文数量的相关性表明，越是位于前列的研究前沿主题的论文数量越高。相比于 ESI 研究前沿的高被引论文聚类，SciVal 研究前沿主题论文数量平均在千篇以上，这导致存在如何准确高效地识别核心论文和解读主题的问题。同时，通过分析多个相关主题识别新兴主题和发现技术转化的潜力和可能路径，都是值得深入探索的问题。

下一步的研究将基于已经发布的研究主题，从 3 个方面开展更多实证分析：①引入 Altmetrics 指标，比较媒体关注与学术影响力在主题探测中的差异和影响；②将显著性主题用于机构和学科在研究前沿上的竞争力分析评价；③开展具体研究领域或研究方向的知识演化图谱和技术转化预测研究。

参考文献：

[1] 吴菲菲，杨梓，黄鲁成. 基于创新性和学科交叉性的研究前沿探测模型——以智能材料领域研究前沿探测为例[J]. 科学学研究，2015，33(1):11 - 20.

[2] GARFIELD E. Citation indexes for science: a new dimension in documentation through association of ideas[J]. Science,1955,122(3159): 108 - 111.

[3] PRICE D J. Networks of scientific papers[J]. Science, 1965, 149(3683):510 - 515.

[4] KUHN T S. The structure of scientific revolutions[M]. Chicago: University of Chicago Press,1970.

[5] GARFIELD E, SHER I H, TORPIE R J. The use of citation data in writing the history of science[M]. Philadelphia: Institute for Scientific Information, 1964.

[6] KESSLE M M. Comparison of the results of bibliographic coupling and analytic subject indexing[J]. American documentation, 1965, 16(3):223 - 233.

[7] SMALL H, GRIFFITH B C. The structure of scientific literatures, I: identifying and graphing specialties[J]. Social studies of science, 1974(4):17 - 40.

[8] SMALL H. Tracking and predicting growth areas in science[J]. Scientometrics, 2006, 68(3):595 - 610.

[9] SMALL H. Visualizing science by citation mapping[J]. Journal of the Association for Information Science & Technology, 1999, 50(9):799-813.
[10] KLAVANS R, BOYACK K W. Quantitative evaluation of large maps of science[J]. Scientometrics, 2006, 68(3):475-499.
[11] 王小梅,邓启平,李国鹏,等. ESI 研究前沿的科学图谱及在纳米领域的应用[J]. 图书情报工作,2017,61(12):106-112.
[12] 周群,周秋菊,冷伏海. 生物科学研究前沿演进时序分析[J]. 中国科学院院刊,2017(4):405-412.
[13] 冷伏海,祝清松. 关键研究路径分析方法优化及应用研究——以量子失协领域为例[J]. 情报科学,2016(4):3-6,12.
[14] BOYACK K W. Using detailed maps of science to identify potential collaborations[J]. Scientometrics, 2009, 79(1):27-44..
[15] KLAVANS R, BOYACK K W. Toward an objective, reliable and accurate method for measuring research leadership[J]. Scientometrics, 2010, 82(3):539-553.
[16] KLAVANS R, BOYACK K W. Using global mapping to create more accurate document-level maps of research fields[J]. Journal of the American Society for Information Science & Technology, 2011, 62(1):1-18.
[17] WALTMAN L, VAN ECK N J. A new methodology for constructing a publication-level classification system of science[J]. Journal of the Association for Information Science & Technology, 2012, 63(12):2378-2392.
[18] BOYACK K W, KLAVANS R. Creation of a highly detailed, dynamic, global model and map of science[J]. Journal of the Association for Information Science & Technology, 2014, 65(4):670-685.
[19] KLAVANS R, BOYACK K W. Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? [J]. Journal of the Association for Information Science & Technology, 2016, 68(4):984-998.
[20] SMALL H, BOYACK K W, KlAVANS R. Identifying emerging topics in science and technology[J]. Research Policy, 2014, 43(8):1450-1467.
[21] 2017 研究前沿[EB/OL]. [2018-03-17]. https://clarivate.com.cn/research_fronts_2017/2017_research_front.pdf.
[22] BOYACK K W, KLAVANS R, SMALL H, et al. Characterizing the emergence of two nanotechnology topics using a contemporaneous global micro-model of science[J]. Journal of Engineering & Technology Management, 2014, 32(32):147-159.
[23] Delving deeper into topic prominence in science[EB/OL]. [2018-03-17]. https://www.elsevier.com/__data/assets/pdf_file/0006/548313/Topic-Prominence-Advanced-Webinar.pdf.
[24] ECK N J V, WALTMAN L, DEKKER R, et al. A comparison of two techniques for bibliometric mapping: multidimensional scaling and VOS[J]. Journal of the American Society for Information Science & Technology, 2010, 61(12):2405-2416.
[25] KLAVANS R, BOYACK K W. Research portfolio analysis and topic prominence[EB/OL]. [2017-11-25]. https://arxiv.org/ftp/arxiv/papers/1709/1709.03453.pdf.
[26] 徐淑姝. 面向卓越性的百分位数指标应用研究[D]. 北京:北京理工大学, 2015.
[27] SIEGEL R L, MILLER K D, JEMAL A. Cancer statistics[J]. Cancer Journal for Clinicians, 2016, 66(1):7-30.

作者贡献说明:
崔宇红:方法研究和论文撰写;
王飒:数据分析;
高晓薇、曹学伟:主题解读;
杨卉:数据提供。

Detecting and Characterizing Research Fronts Topics Based on Global-Micro Model

Cui Yuhong[1]　Wang Sa[1]　Gao Xiaowei[2]　Yang Hui[3]　Cao Xuewei[2]
[1] Beijing Institute of Technology Library, Beijing 100081
[2] National Academy of Innovation Strategy, Beijing 100012
[3] Relx Group Shanghai District, Shanghai 200040

Abstract: [**Purpose/significance**] Accurate judgment of research fronts is the national strategic macro-level demand, and scientometrics is commonly used in the quantitative method of research fronts and topic detection. [**Method/process**] Firstly, literature review is focused on topic detection and research fronts, then concept of the global-micro model and methods in topic creation are introduced in detail, including topic cluster with direct citation, name label with keyword, and selection methodology of topic prominence. It also analyzes nearly 96,000 topics and the top 1% research fronts created by Scival. [**Result/conclusion**] The global-micro model can identify all topics of different fields at the same time, but there are differences in the research fronts between different subjects, which can not equate topic prominence to the importance of simplicity. There is a moderate correlation between the number of topic papers and the topic ranking. Automatically extracted keywords can be named and described the topic in terms of the subject level and uniqueness. The topic evolution is demonstrated by the related research fronts of graphene, which can be used to identify key events and emerging trends.

Keywords: topic detection　research fronts　global-micro model　SciVal　topic prominence